Abstract Title Page
Not included in page count.

Title: Selecting a sample for your experiment: A non-random stratified sampling approach

Authors and Affiliations: Elizabeth Tipton, *Teachers College, Columbia University*

Abstract Body

Background / Context:

Randomized experiments in education research are powerful tools for determining if the effect of a treatment or intervention on a particular population *causes* changes in important educational outcomes. For this reason, experiments are often considered the "gold standard" for education research. Indeed, even when experiments are not possible, state-of-the-art methods aim to approximate an experiment through methods like propensity score matching, regression discontinuity, or instrumental variables (Shadish, Cook, and Campbell, 2002).

While randomized experiments answer important scientific questions about causality, they often do so for populations that are ill defined, making generalizations to policy relevant populations difficult. These generalizations would, of course, be greatly improved if units were randomly selected from a well-defined population into the randomized experiment. While this dual randomization is ideal, it is generally infeasible given the practical difficulties and constraints of sample recruitment (Shadish, Cook, and Campbell, 2002). Since methods for sample selection for generalization *not* based on random sampling are currently not available, generalizations from experiments are oftentimes astatistical and simplistic in nature. This is the case even when generalization is the focus of the experiment, as is the case in IES Goal 4 Scale-up studies.

Recently, new methods have been developed for improving generalizations from completed experiments to well-defined populations of interest (Hedges & O'Muircheartaigh, in press; Stuart, Cole, Bradshaw, & Leaf, 2011; Tipton, under review). This work extends propensity score matching methods commonly used to improve the internal validity of observational studies to the problem of generalization. A key feature of this retrospective work is that it matches units in the experiment to units in a population via a sampling propensity score. This propensity score is defined in relation to important covariates that may explain heterogeneity in treatment effects.

While this recent work carefully develops theory and methods for the retrospective generalization case, the problem of prospective sample selection for experiments is only beginning to be addressed. Recently a method was proposed by Tipton, Sullivan, Hedges, Vaden-Kiernan, Borman, and Caverly (2011) to improve these generalizations in the special case in which the population of interest and the population of eligible units does not overlap. This method divides the generalization population into strata based on a sampling propensity score that matches eligible units to population units. The goal of this method is to select a sample from the population of eligible units that is compositionally similar to the generalization population on a set of key covariates that potentially moderate the treatment effect.

Purpose / Objective / Research Question / Focus of Study:

The purpose of this paper is to develop a more general method for sample recruitment in experiments that is purposive (not random) and that results in a sample that is compositionally similar to the generalization population. This work builds on Tipton et al. (2011) by offering solutions to a larger class of problems than the non-overlapping eligible and population case described above. There are three parts to this approach, which as given below.

Step 1: Populations and samples defined

The first part of this approach requires researchers to agree on who the generalization population is, what eligibility requirements for the study are, and how large the sample needs to be. To this end we define the following three sets:

(1) A generalization population (P) of size N_p must first be well defined; this requires a population-frame such as a census, administrative data system, or probability survey.
(2) The sample (S) size n and requirements for units (e.g. schools, districts) must be determined a priori based using commonly available methods for power analysis.
(3) The population of eligibles (E) of size N_e must be well defined. This population is based on any constraints or inclusion criteria, including power-analysis, financial, or practical. Examples of these constraints include: the power analysis may dictate that only schools with at least 40 students in first-grade should be included; for financial reasons it would be preferable to only include schools close to major metropolitan areas; or for practical and scientific reasons schools in which the curriculum under study is not currently being used. It may be the case that E ≡ P, but this is not required.

Step 2: Covariate selection assumptions

The second part of this approach requires that a set of covariates must be selected for comparing the realized sample S to the population P. The following two assumptions are necessary:

(A1) *Sampling unconfoundedness*. The set of covariates $\mathbf{X}_p = \{X_1,\ldots,X_p\}$ used for comparisons should contain all those that explain variation in the potential treatment effects. Note that covariates that are associated with potential outcomes but which have the same relationship with both potential outcomes do not need to be included in $\mathbf{X}_p$.
(A2) *Eligible unconfoundedness.* If E and P are not identical, then $\mathbf{X}_p$ cannot contain any covariates that define the set of eligibles E. For example, if only schools with at least 40 first graders are eligible for the study, but the population contains schools that also have less than 40 first graders, then A2 means we must assume that the treatment effects do not vary in relation to the size of the first grade class. If A2 cannot be met, then it may be better to redefine P so that it does.

The remainder of the method developed here hinges on meeting these assumptions. This means that these assumptions should be clearly stated when the method is implemented in practice. If A1 is not met, it means that despite the fact that the sample S and population P are balanced on the $\mathbf{X}_p$ covariates, there may be other covariates that explain variation in treatment effects but upon which S and P are not balanced (leading to bias). This assumption should not sway researchers from using this method, however, since current practice often leads to situations in which S and P are at most balanced on one or two categorical variables.

Step 3: Sample selection method

Once the population (P), sample (S), eligibles (E), and covariates ($\mathbf{X}_p$) have been well-defined, this paper develops a method for sample selection. The method we develop is an extension of stratified sampling in the univariate case. In stratified sampling, strata or blocks are defined in relation to a single variable X and the sample n is allocated to each stratum j such that $n_j/n = N_j/N = w_j$. The combined estimator can be written

$$T = \Sigma w_j T_j$$

where there are $j=1...k$ strata, and where for each stratum a stratum specific treatment effect T_j can be estimated. For example, T_j may be the simple difference in means estimator between the treatment and control units. In this approach, proportional allocation is used since it results in a sample that is self-weighting (Lohr, 1999).

The problem that this paper addresses is how to proceed with stratified sampling in the case of multivariate $\mathbf{X}_p$. There are three cases we discuss:

(*Case 1*) $S \subset E \equiv P$: In this case any unit from the population can be selected into the sample.

(*Case 2*) $S \subset E \subset P$: In this case, some units of the population are not eligible to be in the experiment. We assume A2 has been met.

(*Case 3*) $S \subset E \not\subset P$: In this case, the sample units must be selected from eligible units which are not in the population. We assume A2 has been met.

In all cases, the goal is to select as a sample S of n units that are compositionally similar to the N units in the population P. This means that S and P are balanced on the covariates in $\mathbf{X}_p$, where balance is defined using the usual measures (e.g. t-tests, standardized mean differences).

Significance / Novelty of study:

Stratified sampling is sometimes used in scale-up experiments, but in only the most basic form. Typically, the population is stratified on only on one or two covariates, and these covariates are generally categorical in nature. For example, it may be that the sample is selected so that there are both urban and rural schools, or so that there are schools in different regions of the country. No methods for the more complex, multivariate and continuous $\mathbf{X}_p$ case have been developed.

The problem of multivariate stratified sampling has, however, been addressed in the survey sampling literature. This paper borrows theory developed in this literature to develop a method to select a purposive stratified sample for experiments when $\mathbf{X}_p$ is multivariate. This method is general, simple to implement, and builds on existing methodologies found in the data mining, survey sampling, and observational studies literatures.

Statistical, Measurement, or Econometric Model:

In the paper we develop the general steps involved in stratifying the population and the creation of a sample selection plans within each of these strata. For all of the cases 1 - 3, the method goes as follows:

(1) Reduce the problem of matching on $\mathbf{X}_p$ to the problem of matching on Z, where $\mathbf{X}_p$ is p dimensions and Z is 1 or 2 dimensions. We will focus here on the 1-dimension case, but note that Tipton et al. (2011) offers an example using the 2-dimension case.

(2) Define k strata based on Z. Note that these strata should be defined in relation to the population P, not E.

(3) Within each of the k strata, allocate the n units in the sample S so that stratum j receives $w_j = n_j/n = N_j/N$ of the sample. This is proportional allocation and is optimal in terms of precision. For example, is ½ of the population P is in stratum 1, then ½ of the sample should also be in stratum 1. Note that in cases 2 and 3, it may be that $w_j \neq N_{ej}/N_e$, meaning that the eligible units are allocated differently to the strata than the population units are.

(4) Develop a method for recruiting units in each stratum. One method is to rank the units in terms of distance from the "center" of the stratum using a distance measure of some sort (e.g. Euclidian or statistical). Another method is to randomly sample units within the stratum. The

key is to develop a recruitment plan that is flexible in the sense that non-response is accounted for.

(5) Allocate resources for recruitment based on information about the *k* strata. This could include information about the stratum sampling fractions (n_j/N_{ej}), the cost of travel to the average unit in the stratum or a measure of difficulty in terms of recruitment (e.g. based on experience from previous studies).

Finally, note that the method used for the first of these steps – reducing the dimensionality of $\mathbf{X}_p$ from *p* to *1* – will depend on the type of data available and the situation. This problem of dimension reduction has been addressed in the classification and clustering literatures. In this paper, we focus on two of these methods:

(1) *Cluster analysis***.** Cluster analysis begins with *n* units and an $\mathbf{X}_p$ matrix and divides the *n* units into *k* clusters, so that units in the same cluster are more similar to one another than units in other clusters (e.g. Everitt, Landau, Leese, & Stahl, 2011). We propose that this method can be used in Case 1, but also in Case 2 if $N_e \approx N_p$.

(2) *Propensity score matching***.** This is a type of classification or discrimination approach that defines a function (the propensity score) that classifies units into one of two groups (here E or P/E; Rosenbaum & Rubin, 1983). Strata can be defined by partitioning the distribution of the propensity score in the population into *k* equal groups (Cochran, 1968; Rosenbaum & Rubin, 1984). We propose that this method can always be used in Case 3, but is also useful in Case 2 when $N_e/N_p \in (0.10, 0.90)$, which is to say when E and P differ by at least 10%.

Usefulness / Applicability of Method:

In order to demonstrate the usefulness of this sample selection method, we include two examples using data from the Texas AEIS administrative data system. In both examples we assume that the generalization population is the population of middle schools in Texas, that a power analysis has determined that *n=40* schools must be selected, and that $\mathbf{X}_p$ includes 22 school aggregated variables. These examples focus on Cases 1 and 2; note that an example of Case 3 can be found in Tipton et al. (2011).

The first of these is an example of Case 1, where there are no constraints on the sample (and in which all population units are eligible for the study). For this case, we use a cluster analysis approach to divide the population into *k* strata (where $k \leq 40$).

The second of these is an example of Case 2, where there are constraints on the sample. We assume that only schools with at least 200 9^{th} graders are eligible to be in the study, based on the results of a power analysis. For this example, we use two approaches: a cluster analysis approach and a propensity score matching approach.

Conclusions:

This paper provides a general method for purposive stratified sampling on key covariates that potentially explain variability in treatment effects. The goal is for the sample to be selected so that it is compositionally similar to a well-defined policy relevant population. Additionally, this paper offers a framework that focuses discussions around generalization in terms of populations, eligibility, treatment effect heterogeneity, and resource allocation. By addressing these issues before sample selection begins, the likelihood of an experiment leading to a causal *and* generalizable result is greatly increased.

Appendices

Appendix A. References

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*(2), 295-313.

Everett, B.S., Landau, S., Leese, M., & Stahl, D. (2011) Cluster analysis. Wiley Series in Probability and Statistics. John Wiley & Sons: West Sussex, UK.

Hedges, L.V. and O'Muircheartaigh, C.A. (*under review*) Improving generalization from designed experiments.

Lohr, S. (1999) *Sampling: Design and analysis.* Duxbury Press: Pacific Grove, CA.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55. doi: 10.1093/biomet/70.1.41.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*(387), 516-524. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20199225.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston: Houghton-Mifflin.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society, Series A*, Part 2, 369-386.

Tipton, E. (*under review*). Improving the external validity of randomized experiments using propensity score subclassification.

Tipton, E., Sullivan, K., Hedges, L.V., Vaden-Kiernan, M., Borman, G., & Caverly, S. (2011) Designing a sample selection plan to improve generalizations from two scale-up experiments. *Abstracts of papers, Fall Meeting of the Society for Research on Educational Effectiveness*, Washington, D.C.